Solution 1: Bayesian Linear Model

The posterior distribution is obtained by Bayes' rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})}^{\text{likelihood}}\overbrace{q(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}.$$

In the Bayesian linear model we have a Gaussian likelihood: $\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}^\top\boldsymbol{\theta}, \sigma^2\boldsymbol{I}_n)$, i.e.,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) &\propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right] \\ &= \exp\left[-\frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{\sum_{i=1}^n(y^{(i)}-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2}\right]. \end{aligned}$$

Moreover, note that the maximum a posteriori estimate of $\boldsymbol{\theta}$, which is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})$$

can also be defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})\right),$$

since log is a monotonically increasing function, so the maximizer is the same.

(a) If the prior distribution is a uniform distribution over the parameter vectors $\boldsymbol{\theta}$, i.e.,

$$q(\boldsymbol{\theta}) \propto 1,$$

then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{\sum_{i=1}^n(y^{(i)}-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2}\right]. \end{aligned}$$

With this,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})\right) \\ &= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^n(y^{(i)}-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} \\ &= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^n(y^{(i)}-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n(y^{(i)}-\boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2, \qquad (2\sigma^2 \text{ is just a constant scaling}) \end{aligned}$$

so the maximum a posteriori estimate coincides with the empirical risk minimizer for the L2-loss (over the linear models).

(b) If we choose a Gaussian distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \qquad \tau > 0,$$

then

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})q(\boldsymbol{\theta}) \\
&\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{1}{2\tau^2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right] \\
&= \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2}\right]
\end{aligned}$$

With this,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\right) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\theta}\|_2^2,
\end{aligned}$$

so the maximum a posteriori estimate coincides for the choice of $\lambda = \frac{\sigma^2}{\tau^2} > 0$ with the regularized empirical risk minimizer for the L2-loss with L2 penalty (over the linear models), i.e., the Ridge regression.

(c) If we choose a Laplace distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^{p}|\boldsymbol{\theta}_i|}{\tau}\right], \qquad \tau > 0,$$

then

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})q(\boldsymbol{\theta}) \\
&\propto \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\sum_{i=1}^{p}|\boldsymbol{\theta}_i|}{\tau}\right] \\
&= \exp\left[-\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_1}{\tau}\right]
\end{aligned}$$

With this,

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\right) \\
&= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_1}{\tau} \\
&= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{\|\boldsymbol{\theta}\|_1}{\tau} \\
&= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}(y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \frac{2\sigma^2}{\tau}\|\boldsymbol{\theta}\|_1,
\end{aligned}$$

so the maximum a posteriori estimate coincides for the specific choice of $\lambda = \frac{2\sigma^2}{\tau}$ with the regularized empirical risk minimizer for the L2-loss with L1 penalty (over the linear models), i.e., the Lasso regression.